PRELIMINARY: COMMENTS VERY WELCOME

Revisions to the Variance Estimation Procedure for the SCF

Arthur B. Kennickell
Senior Economist and Project Director Survey of Consumer Finances
Mail Stop 153, Federal Reserve Board, Washington, DC 20551
Phone: (202) 452-2247
Fax: (202) 452-5295
Email: Arthur.Kennickell@frb.gov
SCF Web Site: http://www.bog.frb.fed.us/pubs/oss/oss2/scfindex.html

October 2000

The views presented in this paper reflect the judgments of the author alone and do not necessarily reflect the opinions of the Board of Governors of the Federal Reserve System or its official staff. The author is grateful, as always, to Fritz Scheuren for discussions. Any errors or follies committed in this paper are the responsibility of the author alone.

The Survey of Consumer Finances (SCF) uses a bootstrap technique for computing estimates of sampling variance. Although bootstrap procedures may not always be the theoretically best option (Sitter, 1992), in surveys with samples as complicated as that for the SCF, there is sometimes no other feasible general alternative. To provide reasonable estimates of sampling variances, bootstrap methods should exploit the important dimensions of variability within the set of completed survey cases, that might have occurred in the selection of the original sample and its implementation in the field. In the SCF application of this approach, sample replicates are selected, and weights are computed for each of these replicates using the standard weighting SCF algorithm (Kennickell and Woodburn, 1999) as if each of the selected replicates were the full set of completed cases.

Over time, the variances estimated for the SCF have been subjected to intensive review, but for many of the estimates made with the survey, it is very difficult to develop a reliable alternative estimate to use for comparison. However for percent distributions, the simple random sampling (SRS) estimator of variance provides a point of reference. Recent work looking at percent distributions of the population over wealth groups using SCF data revealed that the estimated variances for these estimates are implausibly larger than the SRS estimates. This work led to further review of the SCF variance estimation methodology and the proposed revisions presented in this paper.

The first section of this paper gives some background on the survey and discusses the framework used for estimating sampling variances for the survey. The second section presents a set of the estimates that provoked this review and discusses the sources of what may be characterized as excessive estimated variability. The third section proposes a modification of the variance estimation procedure for the SCF. The final section summarizes the paper and offers some thoughts for future research.

I. SCF variance estimation methodology

A. Background on the survey and its sample design

Beginning with 1983, the SCF has been conducted on a triennial basis. The 1989 survey marked a major revision of the methodology for the survey, which has been maintained as constant as possible since then. The SCF is sponsored by the Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division (SOI) of the Internal

Revenue Service. Before 1992, the data for the survey were collected by the Survey Research Center at the University of Michigan, and since that time, the data have been collected by the National Opinion Research Center at the University of Chicago.

The survey is intended to collect detailed information on the finances of U.S. families, and this mission has a strong effect in determining the sample design. Many populations characteristics, such as ownership of credit cards and home mortgages are widely distributed. However, it is also the case the wealth is highly concentrated (Kennickell, 2000), and EPSEM samples will be very unlikely to obtain sufficient cases to support sufficiently robust estimation of many wealth-related characteristics. Moreover, the available evidence suggests strongly that nonresponse is correlated with wealth (Kennickell and McManus, 1993), and estimation that does not have a means of dealing with this problem will produce biased estimates of many statistics.

To address these constraints of the survey, the SCF employs a dual-frame sample design. One part is an national multi-stage area-probability (AP) sample that gives good coverage of the general population (Tourangeau *et al.*, 1993). The second part of the sample is selected as a list from statistical records derived from tax returns by SOI; this sample is designed to over-sample wealthy households (Kennickell 1998b). This list sample provides a large number of observations to support analysis of many characteristics that are strongly influenced by the upper tail of the wealth distribution, and the sample also gives a powerful tool for dealing with nonresponse that is associated with wealth.

The AP sample is selected in stages.[1] At the first stage, the U.S. is divided into geographic groups ranging in size from the very largest metropolitan areas to individual rural counties. Some areas are selected as primary sampling units (PSUs) with probability one; in the most recent AP sample that formed the basis of the 1998 SCF, there were 19 such areas. The remaining areas are stratified by various factors, and PSUs are selected from the strata with probabilities proportional to the populations of the areas. In the sample used for the 1998 SCF, there were 81 non-self-representing PSUs. Within each of the selected PSUs, sub-areas are

[1]There are some differences in the SCF sample designs over time. Where there are differences in detail, the exposition in this paper follows the design of the 1998 survey.

selected using another stratification scheme. From the lowest geographical unit—roughly, the "block" level—individual housing units are selected.

There is only one sense in which there is meaningful dependence between the AP and list samples. In order to limit the cost and management complexity of the survey, the geographic range of the list sample is constrained to the first-stage PSUs selected for the AP sample. As Frankel and Kennickel (1995) have noted, the distribution of wealthy families across the country differs substantially from that of the general population. Thus, using the population-based PSU selections for the AP sample makes the estimates provided by the list sample less efficient than they might be if the areas chosen for the list sample were optimized independently. However, as those authors concluded, taking all of the PSUs together, the coverage of wealthy households is sufficiently good. Given the set of PSUs, the list sample cases are selected from the SOI data stratified by a "wealth index," which is an approximation of the relative wealth of each sample element.

B. A summary of sampling variance estimation procedures in the SCF

As is commonly the case, the sampling variance estimation methodology used for the SCF attempts to mimic the sort of variation that was associated with the actual selection and execution of the survey. The SCF uses a bootstrap procedure to draw 999 replicate samples from the completed sets of AP sample cases and list sample cases, and a weight is calculated for each replicate using the same procedures applied for the full set of observations (see Kennickell and Woodburn, 1999). An estimate of the sampling variability of a given survey estimate is obtained by making the estimate with each replicate (and weight, where appropriate) and then computing the standard deviation of the replicate estimates.

Following the practice of many other surveys, the replicate samples chosen from the completed AP cases take the level of PSU selection as the basic unit of variability for the non-self-representing areas. The original sample was drawn in such a way that the non-self-representing PSUs may be grouped into pseudo-strata. Generally, there are two PSUs per pseudo-stratum, but in some cases there are three such areas. For the bootstrap samples, these PSUs are sampled with replacement from the pseudo-strata up to the number of areas originally selected. For the self-representing areas, comparable pairs of sub-areas serve as the basis of the replicate sample. As input into the analysis weight constructed for each replicate, a post-

stratification-adjusted weight for the AP sample cases alone is computed using the original selection weights, the original number of cases selected in each PSU, and various other post-strata controls including the age of the head of the household and the housing tenure status of households.

Reflecting the common geographic structure of the two parts of the SCF sample, when a given non-self-representing PSU is selected into an AP bootstrap replicate, all list sample cases in that area are included in the corresponding list sample replicate. Because the list sample does not employ any of the geographic selection below the PSU level in the AP sample, it is not possible to make use of the sub-areas in the self-representing areas for constructing the bootstrap replicates. In such areas, bootstrap samples are selected by simple random sampling with replacement within the wealth index strata, where the number of cases selected is equal to the number originally interviewed in those areas. In parallel with the AP replicate sample weights, non-response adjusted post-stratification weights are also computed for the list sample replicates using frame information, including wealth index stratum totals, a measure of financial income, and geographic information.

There are at least two ways that one might use the two parts of the sample for joint estimation. First, one might make estimates with each part separately, and then use information on relative sample sizes and other information related to differential estimation bias and efficiency to pool the separate estimates. This approach raises several problems. Such estimation would require information on the sample design to be included in the public version of the SCF dataset, but such information cannot be released for confidentiality reasons. There are also response and frame problems in both samples that would require complex adjustments in order to avoid bias. Moreover, even if the necessary information for such calculations could be given to users, such an exercise would have to be performed for every estimate. For all of these reasons, this approach to pooled estimation is not followed in the SCF.

A second approach is to develop a combined weight for the two parts of the sample. In this case, there is no need to release detailed sample data, and the complex analysis of the relative strengths of the two samples needs to be done only once. The most straightforward way of combining the weights would be to compute for a given case *i*

$$W_i = \pi_i^{-1} = [\pi_i^{AP} + \pi_i^{LS} - \pi_i^{AP}\pi_i^{LS}]^{-1}$$

where W represents the combined weight, which is equal to the inverse of π_i, the joint probability of observation (the product of the probability of selection and the probability of response) under either sample. Unfortunately, the probability of response is not clearly known for either sample. Using the methods referred to above, it is possible to make some steps toward an adjusted weight for each sample independently. But computing such weights for cases under the alternative sample raises far too many complications and the need for far too many assumptions for this approach to be useful.

As an alternative means of computing combined weights for the two samples, the SCF employs a post-stratification technique using sample-based estimates of the number of households in various post-strata defined in terms of their gross assets. In general, the list sample is assumed to represent better the top end of the wealth distribution than does the AP sample, and it also offers some means of adjusting for differential nonresponse in that wealth region. In contrast, the AP sample does a much better job than the list sample of representing the lower end of the distribution—indeed, the list sample does not contain any households that did not file a tax return. In the range between these extremes of wealth, both samples are informative. Within each post-stratum, the final separate sample weights are multiplied by a factor that accounts for the relative contribution of each sample to the estimate of the number of observations in the post-stratum.[2] For the top wealth groups, the post-stratum totals are forced to the estimate derived purely from the list sample. For the remaining post-strata, the overall total is set by the difference between an overall population estimate derived from the March Current Population Survey and the total for the top groups. Finally, the merged weight is further

---

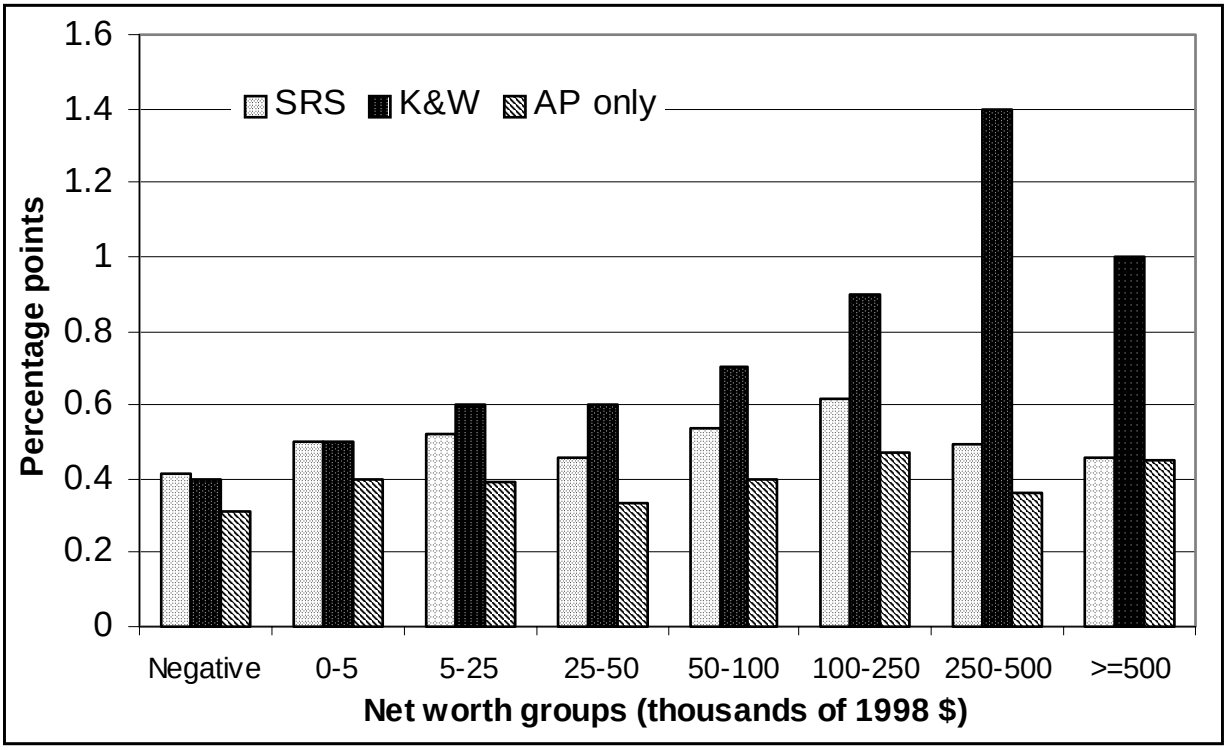
[2]In gross asset post-stratum i, let N_{ia} = weighted number of AP cases, N_{il} = weighted number of list cases, n_{ia} = the unweighted number of AP cases, n_{il} = the unweighted number of list cases, and let $R_{is} = (n_{is}/N_{is})/[(n_{ia}/N_{ia}) + (n_{il}/N_{il})]$ for s={a,l}. Then for case j from sample s in post-stratum i, $COMBINED\_WGT_j = R_{ia} * AP\_WGT_j + R_{il} * LIST\_WGT_j$, where AP_WGT_j is the nonresponse-adjusted AP weight (equal to zero for list cases), and $LIST\_WGT_j$ is the nonresponse-adjusted list weight (equal to zero for AP cases). If the weighted number of AP and list cases were the same in each post-stratum (i.e., $N_{ia}=N_{il}$), then the rescaling would reduce to a simple proportional adjustment based on the relative sample counts.

adjusted to ensure alignment of key populations characteristics, such as the age distribution. This weight, computed for all the bootstrap replicates, serves as the basis for many types of sampling variance estimation in the SCF

II. Possible sources of inflation of the estimated sampling variances

Figure 1 presents estimates of the standard error due to sampling for estimates of the percent of families with net worth in a range of groups in 1998. Such estimates are given for what would be expected under simple random sampling with the same number of observations as

Figure 1: Estimates of sampling variance of proportion of families in various net worth groups; simple random sampling estimate, standard SCF estimates, and estimate using area-probability sample alone (adjusted for sample size difference); 1998 SCF.



the full 1998 SCF, for the actual sample under the variance estimation methodology outlined in the previous section, and for the AP sample alone.[3]

From the figure, it is quite clear that estimates made using only the AP sample and the adjusted replicate weights for that sample, have lower standard errors than would be the case under an equivalent sized simple random sample (that is, the design effect is estimated to be less than one). But when the list sample is included and the merged weights are used, the estimated standard errors are larger than the corresponding estimates for the AP sample in all instances, and larger than the simple random sampling estimates by an increasingly larger margin with increasing levels of wealth. The idea that one might actually suffer a substantial loss in precision from the inclusion of the list sample cases appears questionable, and the size of the loss seems implausible. A similar pattern of large variability in estimates using the full sample relative to estimates using the AP sample alone is sustained for many other estimates that are influenced by the upper part of the wealth distribution (for example, estimates of wealth concentration). The need to resolve this problem was the motivation for the investigation that led to the work reported in this paper.

The calculation of the replicate weights offers many places where a distortion might be introduced into the variance estimates, even if the "main" analysis weight were not affected. Intensive review of the software did not reveal any errors at the level of the implementation of the weighting algorithm. In conducting a number of experiments to extract the contribution to estimated variance of the individual adjustments within the algorithm, it quickly became clear that a large inflation of estimated variances occurs when the AP and list samples are assembled using the post-stratification technique described above, rather than at a more elemental levels of the calculation. If this inflation is inappropriately large, the fault lies in a conceptual error in the weighting design, the use of input data that are corrupted in some sense, or a combination of the two.

---

[3]The AP estimate were computed using the separate bootstrap replicate weights for that sample The AP cases comprises about two-thirds the observation in the full survey. To put the estimates on an approximately comparable basis, the standard error estimates for the AP sample have been reduced by the square root of the ratio of the number of AP observations to the number of observations in both samples.

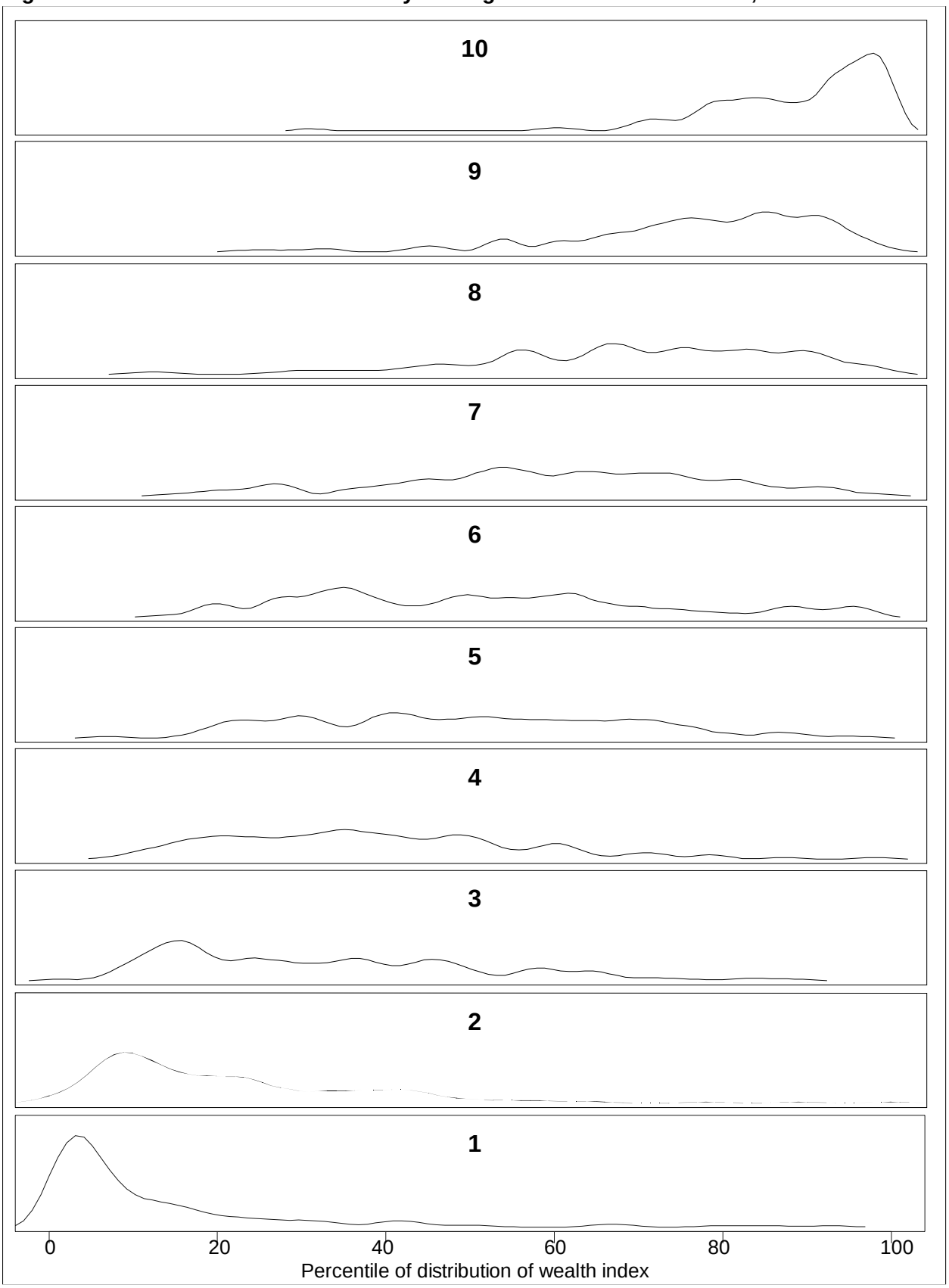
One could argue for alternative weight specifications (indeed, such feedback is welcome), but in designing the weights reported in Kennickell and Woodburn (1999), those authors developed compelling reasons to support the current design. On the bootstrap selection side, there are several possible sources of problems, of which two are potentially important enough to report here. First, as noted earlier in this paper, the grouped PSUs in the non-self-representing areas are much farther from being as well-balanced in terms of numbers of wealthy households than they are in terms of overall numbers of households of all types. Considering only cases in top four list sample strata—the wealthiest cases—the mean ratio of the number of such observations in the smaller of a pair (or the second largest for groups of three) to the number in the largest in a group for the non-self-representing areas is about 66 percent, and the standard deviation of the estimate is about 30 percentage points. However, only about 45 percent of the observations in the list sample are in non-self-representing areas, and adding specific controls to force the number of cases selected by stratum to be constant across the replicates does not alter the estimated variances very much. Second, the number of actual (unduplicated) observations in the list sample varies considerably. For the 1998 SCF, the mean number of list sample observations selected into a bootstrap replicate was 664 out of an actual sample size of 1,496—or only 44 percent of the number of completed interviews in this sample. If one were comparing simple random samples, the expected standard error due to sampling for a sample of 664 would be about 50 percent larger than would be the case for a sample of 1,496, suggesting a possible basis for addressing the variability of the post-stratum estimates. A related issue is the amount of variability allowed across PSUs in the distribution of observations. The survey contract called for a minimum number of observations in each of the list sample strata, and great pains are taken to try, to the degree feasible, to avoid concentrated geographic areas of nonresponse. For the non-self-representing areas, such variation in the replicate samples follows directly from the sampling within PSU groups. For the self-representing areas, the bootstrap selection randomizes over all such PSUs within strata, and one result is that many replicates contain observations that are far more clustered geographically than would have been permitted in fact.

If the correlation of the wealth index used in creating the list sample strata and other possible post-stratifiers for that sample were very strongly correlated with actual net worth, it

might be that the variability of the gross assets cell totals used in the post-stratification could be correspondingly reduced. Kennickell (1998a) provides detailed information on the relationship between net worth and the original wealth index used in constructing the sampling strata, and figure 2 gives an update for the 1998 SCF of one of the key figures in that paper. The figure divides the list sample into unweighted deciles of net worth (given by the stack of rectangles in the figure, with the highest decile at the top), and within each decile, the figure shows an estimate of the density of the wealth index.[4] If the wealth index and net worth were perfectly correlated, the distributions in the rectangles would be clustered around a diagonal band from the lower left to the upper right. Although there is a notable diagonal clustering, the values of the index also stray far beyond the diagonal in every net worth decile. Thus, although there is clearly power in the original design variables, it is also clear that randome samples with sampling strata can have large variability in their wealth estimates.

[4]The breakpoints of the decile groups (in thousands of 1998 dollars) beginning with the 10th percentile point are 93; 313; 628; 1,240; 2,173; 3,983; 7,294; 14,942; 35,962.

Figure 2: Distribution of wealth index by unweighted deciles of net worth, 1998 SCF.



III. A proposed alternative for variance estimation in the SCF

If one accepts that the variance of the post-stratum estimates used in combining the AP and list sample weights is artificially inflated, as suggested by the results of the last section, one might take one of three approaches: make changes in the structure of the bootstrap replicates, make adjustments to the final estimated standard errors, or make adjustments at the point of the post-stratification procedure. For the first possibility, it is difficult to see how a change in the type of bootstrap sample could have a major effect on the relevant component of the estimated variance without some sort of control on the frequency with which observations are selected for a given replicate and the geographic variability of those selections; clarity in the implementation of the bootstrap procedure at that level and a desire not to suppress other types of variability where no problem is evident both argue against this approach.[5] For the second possibility, it would, in principle, be straightforward to make an adjustment to the final standard error estimates if one had a particular rationale. However, a problem with an adjustment at this stage is that it is hard to think of a rule for segregating the effect of the post-stratification variability that would be appropriate across a very broad range of estimates. For the third possibility, the one proposed here, adjustment at the level of the actual post-stratification targets the intervention directly at the point where the variance inflation occurs. Given the choice of this method, there is a need to determine an appropriate adjustment.

One approach that is simple to implement at the post-stratification step is to pool the information on the size of the post-strata estimated from both the bootstrap list sample replicates and the full sample of list sample cases as given by[6]

$$a*Boostrap\ \ estimate\ +\ (1-a)*Full\ \ sample\ \ estimate,\ a\in(0,1).$$

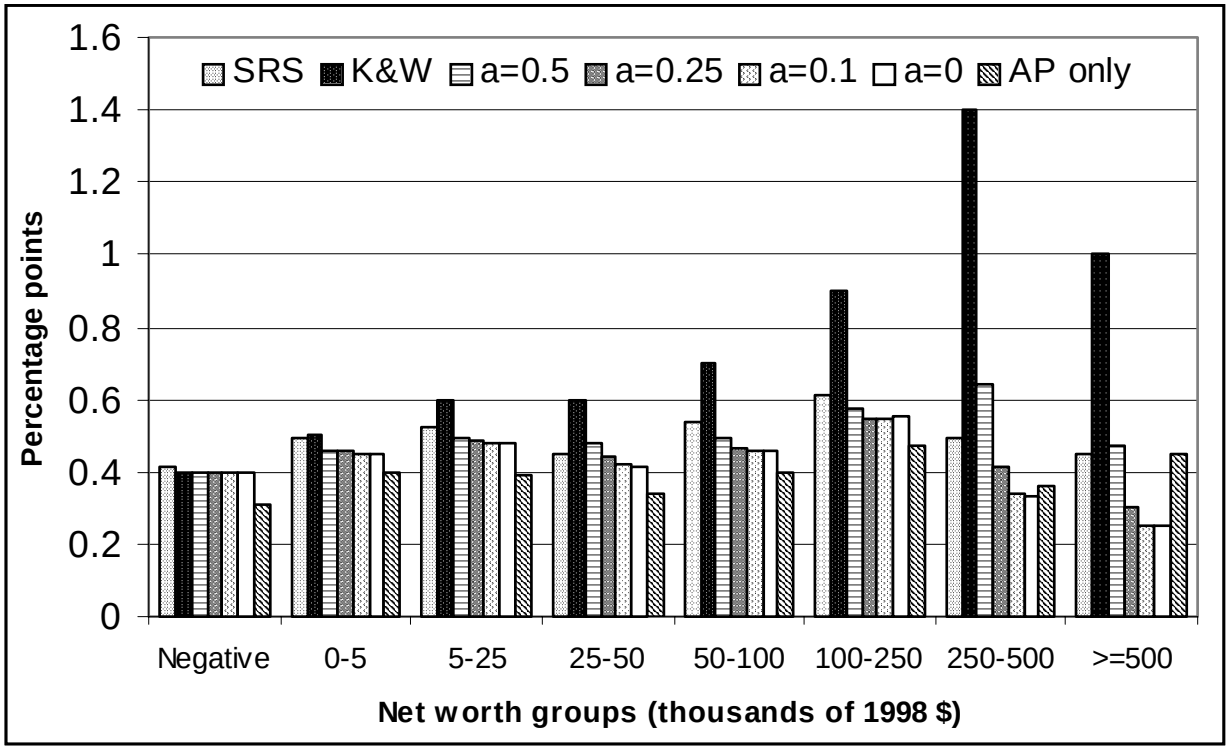
To give a sense of the effects of various degrees of pooling, figure 3 adds a number of pooled estimates to the estimates shown in figure 1. The pooled series shown in the figure are generated by taking values of *a* equal to 1 (the original approach, labeled "K&W"), 0.5, 0.25, 0.1, and 0.

---

[5]Another possibility might be to select the bootstrap samples from the gross asset strata, but such an approach seems to move too far from the original design.

[6]This approach has some motivation in common with "Fay's approach" to balanced repeated replication as reported in Judkins (1990).

Figure 3: Estimates of sampling variance of proportion of families in various net worth groups; simple random sampling estimate, standard SCF estimates, estimates based on various pooled pooled post-strata estimates, and estimate using area-probability sample alone (adjusted for sample size difference); 1998 SCF.



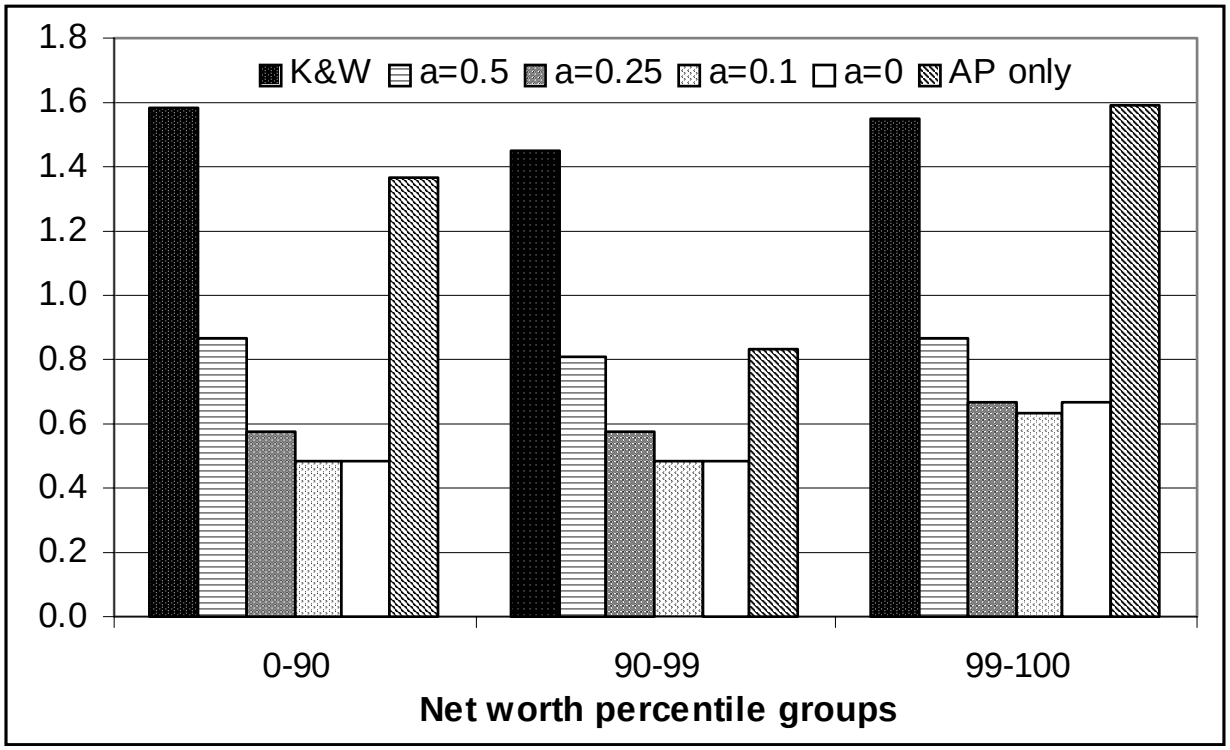
As expected from the nature of the weight construction, the largest effect of the pooling is on the estimated standard errors for the top wealth groups. For the top two groups, the "*a*=0.5" pooling reduces the estimated standard errors by more than half. There are diminishing returns with smaller vales of *a*. The "*a*=0.25" pooling for the top two groups reduced the estimate by about a third of the "*a*=0.5" estimate, and even using only information for the full sample (a=0) reduces the estimate slightly beyond that.

In choosing a value of *a*, two factors provide particular motivation: the reduction in the number of actual cases in the bootstrap samples of the list cases and the geographic variation in the cases in those samples. As noted earlier, the standard error on a percent estimate from a simple random sample with the average number of actual cases in a list sample replicate would be about 50 percent larger than that for a simple random sample with the number of observations in the full list sample. The contribution of the geographic variation to the variability of estimates

is harder to characterize, but casual evidence suggests that it may be substantial. An additional factor to consider is that given that almost 90 percent of the observations in the list sample have net worth of $100,000 or more, it would be very surprising if the final estimates of the distribution of households over the top wealth groups was not less variable than estimates from the AP sample alone. The choice of a=0.25 for pooling seems somewhere in an arguably appropriate range, and it brings the estimated standard error below the adjusted AP estimate for the top percent group and makes them nearly equal for the next highest group.[7]

Of course, one would not want to make a major adjustment to the standard error estimation without considering the effects on the variability of other important estimates. Of a

Figure 4: Estimates of sampling variance of share of total net worth held by families in different groups defined by the percentiles of the wealth distribution; standard SCF estimates, estimates based on various pooled pooled post-strata estimates, and estimate using area-probability sample alone (adjusted for sample size difference); 1998 SCF.



[7]Note that there may well still be comparable problems in variances estimate for statistics computed at sub-national geographic levels. Because pushing the pooled adjustments to a lower geographic level runs the risk of inducing other distortions, that approach is not proposed here.

set of estimates that are strongly affected by the upper tail of the wealth distribution that have been examined, a key set of statistics is the shares of wealth held by different groups of households defined by the percentiles of the wealth distribution. Figure 4 shows standard error estimates under the same range of pooling values used in figure 3. Because there is no analytic formula for the standard error of these estimates under simple random sampling, that comparison is omitted here. The reduction in the standard error of the concentration estimates shows about the same amount of reduction as was the case for the estimates of the percent of families in the top wealth groups. Here the effect is seen in all groups. This result is largely a function of the fact that the denominator (total wealth) is less variable in the pooled estimates.

IV. Conclusion

This paper considers the possibility that the variance estimates for some important estimates using the SCF may be substantially over-estimated. An investigation of the underlying estimation methodology reveals that the variability of some key estimates is dramatically inflated at a point in the replicate weight calculation where the area-probability and list samples in the survey are joined using a post-stratification technique. The post-stratum totals for the number of households in various groups defined in terms of gross asset holdings are derived largely from pooled estimates based on the survey data. Most importantly, the list sample estimates are used to fix the size of the upper tail of the wealth distribution. As it turns out, the list sample estimates of these totals are highly variable under the bootstrap procedure. The relationship between wealth and the stratifying variable in the list sample is not exact, and some "misclassification" results. This type of error leads to variability in the weights within wealth groups. Still, examination of the relationship suggests there is a sufficiently large correlation between wealth and the stratifying variable that it is not credible that the outcome could be a higher level of variability in the post-stratum estimates than under the AP sample. Two factors have strong effects on the estimated variability. First, the distribution of the list sample cases over the geographic areas in the replicate samples is much more varied than would have been allowed in fact. Second, the actual (unduplicated) number of cases in the replicate samples averages less than half the number of cases in the full list sample, the precision of the post-stratum total estimates is strongly affected. To compensate for these factors, the paper argues for

using estimates for the post-strata obtained by pooling information for the replicates with estimates made using the full list sample.

Unfortunately, there is a large arbitrary component of the choice of the pooling factor proposed here. Work should continue in order to develop a foundation for the adjustment—or to overturn it, it that is the appropriate outcome. Work should also be aimed at devising alternative variance estimates that might be used to calibrate the bootstrap estimates. For the 2001 SCF, there are plans to change the selection procedure for the list sample in a way that is likely to sharpen the relationship between wealth and the stratifying variable. In past surveys, a single year of income data has been used to estimate the wealth index used for stratification even though it is known that income can be quite variable over time for reasons only loosely related to current wealth levels. The plan is to extend the estimation to include at least one additional year of income in hopes of smoothing out extraneous income variation.

Bibliography

Frankel, Martin and Arthur B. Kennickell [1995] "Toward the Development of an Optimal Stratification Paradigm for the Survey of Consumer Finances," paper presented at the 1995 Annual Meetings of the American Statistical Association, Orlando, FL.

Judkins, David R. [1990] “Fay’s Method for Variance Estimation,” Journal of Official Statistics, v. 4, no. 3, pp. 223-239.

Kennickell, Arthur B. [1998a] “Using Income Data to Predict Wealth,” paper presented at the Annual Meetings of the Allied Social Science Associations, New York, 1999a.

Kennickell, Arthur B.[1998b] “List Sample Design for the 1998 Survey of Consumer Finances,” working paper, Board of Governors of the Federal Reserve Board, http://www.federalreserve.gov/pubs/oss/oss2/method.html.

Kennickell, Arthur B. [2000b] “An Examination of Changes in the Distribution of Wealth From 1989 to 1998: Evidence from the Survey of Consumer Finances,” working paper, Board of Governors of the Federal Reserve Board, http://www.federalreserve.gov/pubs/oss/oss2/method.html.

Kennickell, Arthur B. and Douglas A. McManus [1993] "Sampling for Household Financial Characteristics Using Frame Information on Past Income," *Proceedings of the Section on Survey Research Methods*, 1993 Annual Meetings of the American Statistical Association, San Francisco, CA.

Kennickell, Arthur B. and R. Louise Woodburn [1999] “Consistent Weight Design for the 1989, 1992, and 1995 SCFs, and the Distribution of Wealth,” *Review of Income and Wealth* (Series 45, number 2), June, pp. 193-215.

Sitter, R.R. [1992] “A Resampling Procedure for Complex Survey Data,” Journal of the American Statistical Association, 87(419), pp. 755-65.

Tourangeau, Roger, Robert A. Johnson, Jiahe Qian, Hee-Choon Shin, and Martin R. Frankel [1993] “Selection of NORC’s 1990 National Sample,” working paper, National Opinion Research Center at the University of Chicago, Chicago, IL.